S-AMP: Approximate Message Passing for General Matrix Ensembles

Cakmak, Burak; Winther, Ole; Fleury, Bernard Henri

Published in:
Information Theory Workshop (ITW), 2014 IEEE

DOI (link to publication from Publisher):
10.1109/ITW.2014.6970819

Publication date:
2014

Document Version
Early version, also known as pre-print

S-AMP: Approximate Message Passing for General Matrix Ensembles

Burak Çakmak
Department of Electronic Systems
Aalborg University
9220 Aalborg, Denmark
Email: buc@es.aau.dk

Ole Winther
DTU Compute
Technical University of Denmark
2800 Lyngby, Denmark
Email: olwi@dtu.dk

Bernard H. Fleury
Department of Electronic Systems
Aalborg University
9220 Aalborg, Denmark
Email: fleury@es.aau.dk

***Abstract*—We propose a novel iterative estimation algorithm for linear observation models called S-AMP. The fixed points of S-AMP are the stationary points of the exact Gibbs free energy under a set of (first- and second-) moment consistency constraints in the large system limit. S-AMP extends the approximate message-passing (AMP) algorithm to general matrix ensembles with a well-defined large system size limit. The generalization is based on the S-transform (in free probability) of the spectrum of the measurement matrix. Furthermore, we show that the optimality of S-AMP follows directly from its design rather than from solving a separate optimization problem as done for AMP.**

***Index Terms*—Variational inference; free energy optimization; approximate message passing; S-transform in free probability**

I. Introduction

Consider an $N \times K$ linear observation model described by

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w} \tag{1}$$

where $\boldsymbol{A} \in \mathbb{R}^{N\times K}$, $\boldsymbol{x} \in \mathbb{R}^{K\times 1}$, $\boldsymbol{y} \in \mathbb{R}^{N\times 1}$, and $\boldsymbol{w} \in \mathbb{R}^{N\times 1}$ are the measurement matrix, the vector to be recovered, the measurement vector, and a white Gaussian noise vector, respectively. The entries of $\boldsymbol{w}$ have zero mean and variance σ_w^2. In [1] the authors propose a recovery scheme for $\boldsymbol{x}$, given $\boldsymbol{A}$ and $\boldsymbol{y}$, called Approximate Message Passing (AMP), which starting from an initial guess $\boldsymbol{\mu}^0 = \boldsymbol{0}$, proceeds iteratively as

$$\boldsymbol{\mu}^{t+1} = \eta_t\left(\boldsymbol{A}^\dagger \boldsymbol{z}^t + \boldsymbol{\mu}^t\right) \tag{2}$$

$$\boldsymbol{z}^t = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}^t + \frac{1}{\alpha}\left\langle \eta'_{t-1}(\boldsymbol{A}^\dagger \boldsymbol{z}^{t-1} + \boldsymbol{\mu}^{t-1})\right\rangle \boldsymbol{z}^{t-1}. \tag{3}$$

The scalar functions η_t, $t \geq 0$, in (2) are obtained by applying an additional optimization procedure based upon the so-called state evolution formula for the underlying measurement matrix ensemble [2]. In (3), $\eta'_t(x) = d\eta_t(x)/dx$, $t \geq 0$. Moreover for a vector $\boldsymbol{u} \triangleq (u_1, \ldots, u_K)$, $\langle \boldsymbol{u}\rangle \triangleq \sum_{k=1}^K u_k/K$ and $\alpha \triangleq N/K$. The vectors $\boldsymbol{\mu}^t$ and $\boldsymbol{z}^t$ are referred to as the current estimate of $\boldsymbol{x}$ and the corresponding residual, respectively. Finally $(\cdot)^\dagger$ denotes transposition.

AMP has two appealing properties. Firstly, when the entries of $\boldsymbol{A}$ are independent identically distributed (iid) Gaussian with zero mean and variance $1/N$, AMP yields the minimum mean square error (MMSE) estimator in the large system limit [2]. Secondly, AMP includes a so-called Onsager reaction term, i.e, $\alpha^{-1}\left\langle \eta'_{t-1}(\cdot)\right\rangle \boldsymbol{z}^{t-1}$ in (3), that corrects the naive mean field approximation. In statistical physics such a technique is known as the Thouless-Anderson-Palmer (TAP) correction [3].

The adaptive TAP (ADATAP) mean field theory was introduced in [4]. In ADATAP the form of Onsager reaction term depends on the measurement matrix, see [4, Eq. (20) & (51)]. Indeed, a connection between ADATAP and AMP has been recently realized in [5]. The connection is based on some approximations of the Gibbs free energy, which are derived using the replica method, see [5, Eq. (10) & (11)] and the references therein.

Inference techniques based on the free energy optimization have become popular in the literature of information theory [6], [7] and in machine learning e.g. [8], [9]. The important results exploited in this contribution is that the fixed points of belief propagation (BP) and expectation propagation (EP) are the stationary points of the Bethe Free energy (BFE) under a set of marginalization consistency constraints [6] and moment consistency constraints [8], respectively.

The conventional approximate message passing methods presented in the literature are based on a Gaussian approximation of loopy BP on a dense graph, [10]–[13]. By contrast, the method presented in this paper is based on probabilistic inference on a tree graph. Specifically we consider an exact Gibbs free energy formulation (i.e. a BFE formulation on a tree probabilistic graph) under first and second-moment consistency constraints. In particular we propose a novel algorithm whose fixed points are the stationary points of the constrained Gibbs free energy in the large system limit. This algorithm – we coin it S-AMP – executes the following iteration steps:

$$\boldsymbol{\mu}^{t+1} = \eta_t\left(\boldsymbol{A}^\dagger \boldsymbol{z}^t + \boldsymbol{\mu}^t\right) \tag{4}$$

$$\boldsymbol{z}^t = \boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}^t + \left(1 - \frac{1}{\mathrm{s}_{\boldsymbol{A}}^{t-1}}\right)\boldsymbol{z}^{t-1} \tag{5}$$

$$\mathrm{s}_{\boldsymbol{A}}^{t-1} \triangleq \mathrm{S}_{\boldsymbol{A}}\left(-\left\langle \eta'_{t-1}(\boldsymbol{A}^\dagger \boldsymbol{z}^{t-1} + \boldsymbol{\mu}^{t-1})\right\rangle\right) \tag{6}$$

with $\mathrm{S}_{\boldsymbol{A}}$ denoting the S-transform of the asymptotic eigenvalue distribution (AED) of $\boldsymbol{A}^\dagger\boldsymbol{A}$ (see, e.g. [14]).

To show that AMP is a special case of S-AMP, let the entries of $\boldsymbol{A}$ be iid with zero mean and variance $1/N$. Then, we have $\mathrm{S}_{\boldsymbol{A}}(\omega) = 1/(1+\omega/\alpha)$ [14, Eq. (2.87)]. Inserting this expression in (6) we obtain the iteration steps (2)-(3) of AMP.

978-1-4799-5999-0/14/$31.00 ©2014 IEEE

Notation: The entries of the $N \times K$ matrix $\boldsymbol{X}$ are denoted by X_{nk}, $n \in \mathcal{N} \triangleq \{1,\ldots,N\}$ and $k \in \mathcal{K} \triangleq \{1,\ldots,K\}$. The entries of a vector $\boldsymbol{u} \in \mathbb{R}^{K\times 1}$ are indicated by u_k. The Gaussian probability density function (pdf) with mean $\boldsymbol{\mu}$ and the covariance $\boldsymbol{\Sigma}$ is denoted by $N(\cdot|\boldsymbol{\mu},\boldsymbol{\Sigma})$. Throughout the paper we assume that $\boldsymbol{A}^\dagger\boldsymbol{A}$ has almost surely an AED as $N,K \to \infty$ while $\alpha = N/K$ fixed.

II. Gibbs Free Energy with Moment Constraints

Consider the linear observation model (1). For Bayesian inference, we assign a prior $p_k(x_k)$ for all $k \in \mathcal{K}$. Hence the joint posterior pdf can be written as

$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{1}{Z} p(\boldsymbol{y}|\boldsymbol{x}) \prod_{k\in\mathcal{K}} p_k(x_k) \tag{7}$$

with $p(\boldsymbol{y}|\boldsymbol{x})$ and Z denoting the likelihood specified by (1) and a normalization constant, respectively. The factor graph representation of (7) is a tree. Thus the BFE of (7) is equal to its Gibbs free energy [6, Theorem 3], which is given by

$$\begin{aligned}\mathrm{G}(\{b_k, b_N, \tilde{b}_k\}) \triangleq &- \sum_{k\in\mathcal{K}} \int b_k(x_k) \log b_k(x_k) dx_k \\ - \int b_N(\boldsymbol{x}) \log \frac{p(\boldsymbol{y}|\boldsymbol{x})}{b_N(\boldsymbol{x})} d\boldsymbol{x} &- \sum_{k\in\mathcal{K}} \int \tilde{b}_k(x_k) \log \frac{p_k(x_k)}{\tilde{b}_k(x_k)} dx_k.\end{aligned} \tag{8}$$

In this expression, b_N and $\tilde{b}_k$, $k \in \mathcal{K}$, denote the beliefs of the factors in (7), while b_k, $k \in \mathcal{K}$, denote the beliefs of the unknown variables in (7).

When we define a Lagrangian for (8) that accounts for the set of marginalization consistency constrains, then at its stationary point, the belief $b_k(x_k)$ is equal to $p(x_k|\boldsymbol{y})$ for all $k \in \mathcal{K}$ [6]. Instead we consider the Gibbs energy formulation with a set of moment consistency constraints. Specifically, following the arguments of [8] we define the Lagrangian

$$\begin{aligned}\mathcal{L}(\{b_k, b_N, \tilde{b}_k\}) \triangleq &\ \mathrm{G}(\{b_k, b_N, \tilde{b}_k\}) + \mathcal{Z} \\ &- \sum_{k\in\mathcal{K}} \bar{\boldsymbol{\nu}}_k^\dagger \int \boldsymbol{\phi}(x_k) \left\{b_N(\boldsymbol{x}) - b_k(x_k)\right\} d\boldsymbol{x} \\ &- \sum_{k\in\mathcal{K}} \boldsymbol{\nu}_k^\dagger \int \boldsymbol{\phi}(x_k) \left\{\tilde{b}_k(x_k) - b_k(x_k)\right\} dx_k.\end{aligned} \tag{9}$$

The term $\mathcal{Z}$ accounts for the normalization constraints for the beliefs:

$$\begin{aligned}\mathcal{Z} \triangleq &-\beta_N \left(1 - \int b_N(\boldsymbol{x}) d\boldsymbol{x}\right) \\ &- \sum_{k\in\mathcal{K}} \beta_k \left(1 - \int b_k(x_k) dx_k\right) - \tilde{\beta}_k \left(1 - \int \tilde{b}_k(x_k)\right)\end{aligned}$$

with $\beta_N, \beta_k, \tilde{\beta}_k, k \in \mathcal{K}$ denoting the associated Lagrange multipliers. We consider constraints on the mean and variance, i.e. $\boldsymbol{\phi}(x_k) = (x_k, x_k^2)$. For convenience we write the Lagrangian multipliers explicitly appearing in (9) in the form

$$\boldsymbol{\nu}_k \triangleq \left[\gamma_k, -\frac{\lambda_k}{2}\right]^\dagger, \quad \bar{\boldsymbol{\nu}}_k \triangleq \left[\bar{\gamma}_k, -\frac{\bar{\lambda}_k}{2}\right]^\dagger, \quad k \in \mathcal{K}.$$

We formulate the estimation procedure for $x_k, k \in \mathcal{K}$, as

$$\mu_k \triangleq \int x_k b_k^\star(x_k) dx_k, \tag{10}$$

where $b_k^\star(x_k)$ represents $b_k(x_k)$ at a stationary point of (9).

A. Stationary Points of the Lagrangian (9)

For the sake of notational compactness we define

$$\boldsymbol{J} \triangleq \frac{1}{\sigma_w^2}\boldsymbol{A}^\dagger\boldsymbol{A}, \qquad \boldsymbol{\theta} \triangleq \frac{1}{\sigma_w^2}\boldsymbol{A}^\dagger\boldsymbol{y} \tag{11}$$

$$\boldsymbol{\Sigma} \triangleq (\boldsymbol{J} + \bar{\boldsymbol{\Lambda}})^{-1}, \quad \boldsymbol{\mu} \triangleq \boldsymbol{\Sigma}(\boldsymbol{\theta} + \bar{\boldsymbol{\gamma}}). \tag{12}$$

In (12) we have introduced the $K \times K$ diagonal matrix $\bar{\boldsymbol{\Lambda}}$ and the $K \times 1$ vector $\bar{\boldsymbol{\gamma}}$ whose entries are respectively $\bar{\Lambda}_{kk} = \bar{\lambda}_k$ and $\bar{\gamma}_k$, $k \in \mathcal{K}$. The stationary points of the Lagrangian (9) are obtained to be of the form

$$\tilde{b}_k^\star(x_k) = \frac{1}{\tilde{Z}_k} p_k(x_k) \exp(\boldsymbol{\nu}_k^\dagger \boldsymbol{\phi}(x_k)), \quad k \in \mathcal{K} \tag{13}$$

$$b_N^\star(\boldsymbol{x}) = N(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \tag{14}$$

$$b_k^\star(x_k) = \frac{1}{Z_k} \exp((\boldsymbol{\nu}_k + \bar{\boldsymbol{\nu}}_k)^\dagger \boldsymbol{\phi}(x_k)), \quad k \in \mathcal{K} \tag{15}$$

with $\tilde{Z}_k$ and Z_k denoting the normalization constants for the beliefs in (13) and (15), respectively. At this stage it is convenient to define $\kappa_k \triangleq \frac{\gamma_k}{\lambda_k}$, $k \in \mathcal{K}$. With this definition we can rewrite the belief (13) in the form

$$\tilde{b}_k^\star(x_k) = \frac{1}{Z(\kappa_k,\lambda_k)} p_k(x_k) N(x_k|\kappa_k, 1/\lambda_k), \quad k \in \mathcal{K}. \tag{16}$$

Furthermore we define for any $k \in \mathcal{K}$

$$\eta(\kappa_k;\lambda_k) \triangleq \kappa_k + \frac{1}{\lambda_k} \frac{\partial \log Z(\kappa_k,\lambda_k)}{\partial \kappa_k}, \tag{17}$$

$$\eta'(\kappa_k;\lambda_k) \triangleq \frac{\partial \eta(\kappa_k;\lambda_k)}{\partial \kappa_k}. \tag{18}$$

It is shown in [12, Eq. (31)-(35)] that $\eta(\kappa_k;\lambda_k)$ and $\eta'(\kappa_k;\lambda_k)/\lambda_k$ give respectively the mean and the variance of the belief (16). With these definitions, the identities resulting from the moment consistency constraints are given by

$$b_k^\star(x_k) = N(x_k|\mu_k,\Sigma_{kk}) \quad k \in \mathcal{K} \tag{19}$$

$$\frac{\lambda_k}{\eta'(\kappa_k;\lambda_k)} = \lambda_k + \bar{\lambda}_k, \quad k \in \mathcal{K} \tag{20}$$

$$\frac{\lambda_k \eta(\kappa_k;\lambda_k)}{\eta'(\kappa_k;\lambda_k)} = \gamma_k + \bar{\gamma}_k, \quad k \in \mathcal{K}. \tag{21}$$

We now derive a simple expression for (10). By making use of the identities in (15) and (19), we write first

$$\gamma_k = \frac{\mu_k}{\Sigma_{kk}} - \bar{\gamma}_k, \quad \lambda_k = \frac{1}{\Sigma_{kk}} - \bar{\lambda}_k, \quad k \in \mathcal{K}. \tag{22}$$

Furthermore by the definitions in (12) we have

$$\bar{\boldsymbol{\gamma}} = -\boldsymbol{\theta} + (\boldsymbol{J} + \bar{\boldsymbol{\Lambda}})\boldsymbol{\mu}. \tag{23}$$

Let us introduce the $K \times K$ diagonal matrix $\mathbf{\Lambda}$ and the $K \times 1$ vector $\boldsymbol{\gamma}$ whose entries are respectively $\Lambda_{kk} = \lambda_k$ and γ_k, $k \in \mathcal{K}$. Then, making use of the identity in (22) we can write

$$\boldsymbol{\gamma} = \boldsymbol{\theta} - (\boldsymbol{J} + \bar{\mathbf{\Lambda}})\boldsymbol{\mu} + \mathrm{diag}(\mathbf{\Sigma})^{-1}\boldsymbol{\mu} \tag{24}$$

$$= \boldsymbol{\theta} - \boldsymbol{J}\boldsymbol{\mu} + \mathbf{\Lambda}\boldsymbol{\mu} = \frac{1}{\sigma_w^2}\boldsymbol{A}^\dagger(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\mu}) + \mathbf{\Lambda}\boldsymbol{\mu} \tag{25}$$

where $\mathrm{diag}(\mathbf{\Sigma})$ is the $K \times K$ diagonal matrix with $\mathrm{diag}(\mathbf{\Sigma})_{kk} = \Sigma_{kk}$, $k \in \mathcal{K}$. Then, by invoking the identities (19) and (21) we arrive at the sought explicit form for (10):

$$\mu_k = \eta(\kappa_k; \lambda_k) \tag{26}$$

$$\kappa_k = \frac{1}{\lambda_k \sigma_w^2} \sum_{n \in \mathcal{N}} A_{nk} \left(y_n - \sum_{l \in \mathcal{K}} A_{nl}\mu_l \right) + \mu_k \tag{27}$$

$$\lambda_k = \frac{1}{\Sigma_{kk}} - \bar{\lambda}_k, \quad \bar{\lambda}_k = \frac{\lambda_k}{\eta'(\kappa_k; \lambda_k)} - \lambda_k, \tag{28}$$

$k \in \mathcal{K}$. As a matter of fact equations (26)–(28) coincide with the fixed-point equations of ADATAP that are obtained by applying the *cavity approach-new* [4] in statistical physics, see [4, Eq. (20), (25) and (26)].

The step in (28) requires a matrix inversion, which is desirable to avoid in order to keep the complexity of fixed-point algorithms devised from (26)–(28) low. In [4] the authors circumvent this complexity problem by using the so-called self-averaging method [4, Section 3.1] in the large system limit under the following underlying assumption:

ASSUMPTION 1 *As $N, K \to \infty$ with the ratio $\alpha = N/K$ fixed let $\boldsymbol{A}^\dagger\boldsymbol{A}$ have an AED almost surely and $\mathrm{E}_{\boldsymbol{A}}\{\lambda_k\} - \lambda_k$ converge almost surely to zero for all $k \in \mathcal{K}$ with $\mathrm{E}_{\boldsymbol{A}}$ denoting expectation over random matrix $\boldsymbol{A}$.*

Showing the explicit necessary conditions for the self-averaging property of $\lambda_k, k \in \mathcal{K}$ in the large system requires an extensive study. Due to space limitation this will be not addressed in this work.

THEOREM 1 *[4, Section 3.1] Let the random matrix $\boldsymbol{A}^\dagger\boldsymbol{A}$ full-fill Assumption 1. Furthermore let $\langle \eta'(\boldsymbol{\kappa}; \mathbf{\Lambda}) \rangle \triangleq \frac{1}{K}\sum_{k \in \mathcal{K}} \eta'(\kappa_k; \lambda_k)$. Then, as $N, K \to \infty$ with the ratio $\alpha = N/K$ fixed, for all $k \in \mathcal{K}$ λ_k converges almost surely to the macroscopic quantity λ that is the solution of*

$$\lambda = \frac{1}{\sigma_w^2} \mathrm{R}_{\boldsymbol{A}} \left(-\frac{\langle \eta'(\boldsymbol{\kappa}; \lambda) \rangle}{\sigma_w^2 \lambda} \right) \tag{29}$$

with $\mathrm{R}_{\boldsymbol{A}}$ denoting the R*-transform (see e.g. [15]) of the AED of $\boldsymbol{A}^\dagger\boldsymbol{A}$ and by abuse of notation $\langle \eta'(\boldsymbol{\kappa}; \lambda) \rangle$ representing $\lim_{K \to \infty} \langle \eta'(\boldsymbol{\kappa}; \mathbf{\Lambda}) \rangle$.*

In [4] the authors show that ADATAP under the self averaging analysis above is consistent with the replica symmetry ansatz. In other words once the replica symmetry ansatz of the underlying system breaks down that spoils the optimality the fixed-point identities (26), (27) and (29).

Making use of the relation between the R-transform and the S-transform [15, Table 6] in (29) we obtain the following corollary.

COROLLARY 1 *Let the random matrix $\boldsymbol{A}$ be defined as in Theorem 1. Then, we have in the large system limit*

$$\lambda = \frac{1}{\sigma_w^2 \mathrm{S}_{\boldsymbol{A}}\left(-\langle \eta'(\boldsymbol{\kappa}, \lambda) \rangle\right)} \tag{30}$$

with $\mathrm{S}_{\boldsymbol{A}}$ denoting the S*-transform of the AED of $\boldsymbol{A}^\dagger\boldsymbol{A}$.*

III. FIXED-POINT ALGORITHMS

In this section we use the stationary point equations obtained in the previous section to devise three fixed point iterative algorithms. Firstly we will present the classical EP scheme for (1) [16] and the ADATAP scheme [4]. Secondly we derive the S-AMP algorithm mentioned in the introduction.

All three recovery schemes have the following update step in common, which results by time-indexing the first identity in (12):

$$\mathbf{\Sigma}^t = (\boldsymbol{J} + \bar{\mathbf{\Lambda}}^t)^{-1}. \tag{31}$$

Since only one element of $\bar{\mathbf{\Lambda}}^t$ is updated in each iteration the matrix inversion lemma can be applied to reduce the complexity of this step to $O(K^2)$, e.g. see [9, Eq. (37)]. This makes (31) suitable for applications with moderately large dimensions of $\boldsymbol{A}$.

A. EP and ADATAP

In the following we present the compact form of the EP scheme for (1) (e.g. see [16]) and the ADATAP scheme [4]. First we start with defining update steps common to both algorithms. They follow by merely time indexing (28):

$$\lambda_k^t = \frac{1}{\Sigma_{kk}^t} - \bar{\lambda}_k^t, \quad \bar{\lambda}_k^t = \frac{\lambda_k^{t-1}}{\eta'(\kappa_k^{t-1}; \lambda_k^{t-1})} - \lambda_k^{t-1}, k \in \mathcal{K} \tag{32}$$

EP [16] updates $\mu_k^t, k \in \mathcal{K}$, based on the second identity in (12), (20) and (27):

$$\mu_k^{t+1} = [\mathbf{\Sigma}^t(\boldsymbol{\theta} + \bar{\boldsymbol{\gamma}}^{t+1})]_k \tag{33}$$

$$\bar{\gamma}_k^{t+1} = \frac{\lambda_k^t \eta(\kappa_k^t; \lambda_k^t)}{\eta'(\kappa_k^t; \lambda_k^t)} - \frac{\mu_k^t}{\Sigma_{kk}^t} \tag{34}$$

$$\kappa_k^t = \frac{1}{\lambda_k^t \sigma_w^2} \sum_{n \in \mathcal{N}} A_{nk} \left(y_n - \sum_{l \in \mathcal{K}} A_{nl}\mu_l^t \right) + \mu_k^t. \tag{35}$$

ADATAP [4] updates $\mu_k^t, k \in \mathcal{K}$, based on the stationary points identities in (26)–(27):

$$\mu_k^{t+1} = \eta(\kappa_k^t; \lambda_k^t) \tag{36}$$

$$\kappa_k^t = \frac{1}{\lambda_k^t \sigma_w^2} \sum_{n \in \mathcal{N}} A_{nk} \left(y_n - \sum_{l \in \mathcal{K}} A_{nl}\mu_l^t \right) + \mu_k^t. \tag{37}$$

Depending on the applications, EP and ADATAP may exhibit a poor convergence behavior, and may even diverge. A procedure to improve the convergence behavior consists in introducing a damping factor, say ϵ, when updating e.g. μ_k^t in (35) and (37) as $(1-\epsilon)\mu_k^t + \epsilon\eta(\kappa_k^t; \lambda_k^t) \to \mu_k^t$. However this approach leads to very slow convergence which might require thousands of iterations, e.g. see [5, Section V]. Regarding more advanced damping schemes we refer the reader to [17].

B. S-AMP

In the sequel we derive a new fixed-point algorithm from the stationary points identities (26)–(28). The algorithm yields S-AMP in the large system limit.

First we return to (27) and define

$$z_{n,k} \triangleq \frac{1}{\lambda_k \sigma_w^2}\left(y_n - \sum_{l\in\mathcal{K}} A_{nl}\mu_l\right), \quad n\in\mathcal{N}, k\in\mathcal{K}. \tag{38}$$

Using this definition in (27) we "devise" the following identity:

$$z_{n,k} = y_n - \sum_{l\in\mathcal{K}} A_{nl}\mu_l + (1-\sigma_w^2\lambda_k) z_{n,k}. \tag{39}$$

Making use of (26), (27) (with definition (38)), and (39) we obtain the new fixed-point algorithm

$$\mu_k^{t+1} = \eta\left(\sum_{n\in\mathcal{N}} A_{nk} z_{n,k}^t + \mu_k^t; \lambda_k^t\right) \tag{40}$$

$$z_{n,k}^t = y_n - \sum_{l\in\mathcal{K}} A_{nl}\mu_l^t + (1-\sigma_w^2\lambda_k^{t-1}) z_{n,k}^{t-1} \tag{41}$$

$k\in\mathcal{K}$, $n\in\mathcal{N}$, where λ_k^t satisfies the system of equations

$$\lambda_k^t = \frac{1}{\Sigma_{kk}^t} - \bar{\lambda}_k^t, \quad \bar{\lambda}_k^t = \frac{\lambda_k^t}{\eta'(\kappa_k^t;\lambda_k^t)} - \lambda_k^t. \tag{42}$$

Like AMP, this scheme includes by design a natural damping factor $(1-\sigma_w^2\lambda_k^{t-1})$ for the contribution $z_{n,k}^{t-1}$. Specifically in this scheme, just like in AMP, we do not need a step-size parameter. However, solving λ_k^t from (42) at each iteration is non-trivial in general. Note that, the scheme in (32) can be considered as an iterative approximation of (42).

Under Assumption 1, by the design of λ_k^t through (42), and from Theorem 1, for all $k\in\mathcal{K}$ λ_k^t converges almost surely to a macroscopic quantity λ^t as $N,K\to\infty$ with the ratio α fixed. Furthermore invoking Corollary 1 the quantity λ^t is the solution of the identity

$$\lambda^t = \frac{1}{\sigma_w^2 \mathrm{S}_{\boldsymbol{A}}\left(-\left\langle \eta_t'(\boldsymbol{\kappa}^t)\right\rangle\right)} \tag{43}$$

where for convenience we define

$$\eta_t(\kappa_k^t) \triangleq \eta(\kappa_k^t;\lambda^t), \quad k\in\mathcal{K}. \tag{44}$$

Consequently we obtain the iteration steps (4)-(6) of S-AMP in their scalar form:

$$\mu_k^{t+1} = \eta_t\left(\sum_{n\in\mathcal{N}} A_{nk} z_n^t + \mu_k^t\right), \quad k\in\mathcal{K} \tag{45}$$

$$z_n^t = y_n - \sum_{l\in\mathcal{K}} A_{nl}\mu_l^t + \left(1 - \frac{1}{\mathrm{s}_{\boldsymbol{A}}^{t-1}}\right) z_n^{t-1}, \quad n\in\mathcal{N}. \tag{46}$$

We note that by definition $\boldsymbol{\kappa}^t = \boldsymbol{A}^\dagger \boldsymbol{z}^t + \boldsymbol{\mu}^t$.

In [2], $\eta_t(\boldsymbol{\kappa}^t)$, $t\geq 0$ in the AMP updates is referred as "an appropriate sequence of non-linear functions". By contrast, in SAMP $\eta_t(\boldsymbol{\kappa}^t)$ results directly from the iterative process designed via the fixed-point equation (43). Note that, λ^t must be solved at each iteration t from this equation. Depending on the prior pdfs, obtaining closed form expression for λ^t is often non-trivial. In fact this shows how S-AMP (or AMP in particular) can be a very advanced estimator as (43) directly relates the asymptotic stationary point identity in (30). In order to better comprehend this aspect, in the following we examine λ^t for the linear Gaussian observations.

1) λ^t for the Linear Gaussian Observation Model: The optimality of AMP for the linear Gaussian models with the zero-mean iid matrix ensemble, was proven in [2, Section 2.1] via a minimization procedure upon the state evolution formula. By contrast we can show the optimality S-AMP for general matrix ensembles from its design.

Consider the linear observation model (1) with the entries of $\boldsymbol{x}$ being iid Gaussian with zero mean and variance one, i.e. $p_k(x_k) = N(x_k|0,1)$, $k\in\mathcal{K}$. Then the asymptotic MMSE of estimating $\boldsymbol{x}$ in (1) reads [14] $\tau_{\boldsymbol{A}}(\sigma_w^2) \triangleq \int (1+\frac{x}{\sigma_w^2})^{-1} \mathrm{d}\mathrm{P}_{\boldsymbol{A}}(x)$ with $\mathrm{P}_{\boldsymbol{A}}$ denoting the AED of $\boldsymbol{A}^\dagger\boldsymbol{A}$. Recall that the fixed points of S-AMP are the stationary points of the Gibbs free energy under the moment consistency constraints in the large system limit. Therefore, for the given a Gaussian prior, S-AMP must be a MMSE estimator in the large system limit. Namely the following relation must be satisfied: $\left\langle \eta_t'(\boldsymbol{\kappa}^t)\right\rangle / \lambda^t = \tau_{\boldsymbol{A}}(\sigma_w^2)$ as $t\to\infty$. We show next that actually *for any* $t\geq 0$, $\left\langle \eta_t'(\boldsymbol{\kappa}^t)\right\rangle / \lambda^t = \tau_{\boldsymbol{A}}(\sigma_w^2)$. First notice that with the choice of the prior we have $\left\langle \eta_t'(\boldsymbol{\kappa}^t)\right\rangle = \lambda^t/(1+\lambda^t)$. Thus, in this case (43) becomes

$$\lambda^t = \frac{1}{\sigma_w^2 \mathrm{S}_{\boldsymbol{A}}\left(-\frac{\lambda^t}{1+\lambda^t}\right)}. \tag{47}$$

The S-transform can be formulated in terms of $\tau_{\boldsymbol{A}}(\sigma_w^2)$ [14, Definition 2.15]. Using this formula we write

$$\frac{1-\tau_{\boldsymbol{A}}(\sigma_w^2)}{\tau_{\boldsymbol{A}}(\sigma_w^2)} = \frac{1}{\sigma_w^2 \mathrm{S}_{\boldsymbol{A}}\left(\tau_{\boldsymbol{A}}(\sigma_w^2)-1\right)}. \tag{48}$$

Thus $\lambda^t = 1/\tau_{\boldsymbol{A}}(\sigma_w^2) - 1$, which confirms the optimality of S-AMP for the linear Gaussian observation model.

IV. A Sub-Optimal Variant of S-AMP

In the previous subsection we derived the explicit expression for λ^t when the prior pdf is Gaussian. However solving λ^t from (43) for other prior pdfs is often non-trivial. A direct approach consists in including an inner loop to solve (43) iteratively at each iteration. That would, however, create an overhead that we would like to avoid. Instead, we propose a sub-optimal scheme for λ_t that does not require any inner loop. Specifically, we approximate λ^t in (43) with λ_s^t that satisfies

$$\lambda_s^t = \frac{1}{\sigma_w^2 \mathrm{S}_{\boldsymbol{A}}\left(-\frac{\lambda_s^t}{\lambda_s^{t-1}}\left\langle \eta_{t-1}'(\boldsymbol{\kappa}^{t-1})\right\rangle\right)}. \tag{49}$$

Note that the fixed points of (49) coincide with (43).

When the entries of $\boldsymbol{A}$ are iid with zero mean and variance $1/N$, the sub-optimal scheme coincides with the classical recursion of AMP used in the literature, see e.g. [12]. In fact, from (49), it is easy to obtain the so-called state-evolution formula [1] for the iid zero-mean matrix ensemble.

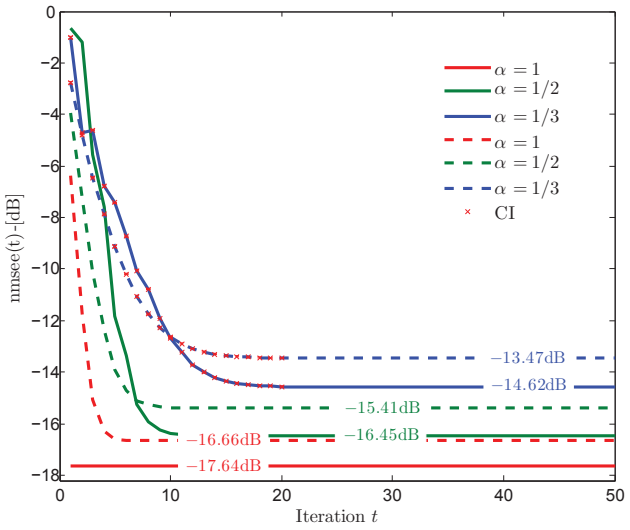

Fig. 1. Performance of the sub-optimal variant of S-AMP implemented for the row orthogonal matrix ensemble (solid curves) and the iid zero-mean Gaussian matrix ensemble (dashed curves). Note that in the latter case the scheme coincides with the classical AMP recursion. The empirical nmsee per iteration is reported versus the number iterations for different selections of α. Confidence intervals (CIs) are also shown for $\alpha = 1/3$. We set $\sigma_w^2 = -20$ dB and $\rho = 0.1$. For each selection of α 2000 trials are performed. The numbers in the plot are the predictions of replica theory [5].

In the sequel we assess the performance of the sub-optimal variant of S-AMP. Due to space limitation we only consider the system model used in [5, Section 5] for Bayesian inference in compressed sensing. Specifically, the prior pdfs are Bernoulli-Gaussian: $p_k(x_k) = (1-\rho)\delta(x_k) + \rho N(x_k|0,1)$, $k \in \mathcal{K}$ with $\rho \in (0,1)$. We refer the reader to [12, Eq. (67) & (68)] for the closed-form expressions of $\eta_t(\cdot)$ and $\eta_t'(\cdot)$ in this case.

We apply the sub-optimal variant of S-AMP in two scenarios: i) the random row-orthogonal measurement matrix ensemble, i.e. $\boldsymbol{A} = \alpha^{-\frac{1}{2}}\boldsymbol{P}_\alpha \boldsymbol{O}, \alpha \leq 1$, where $\boldsymbol{P}_\alpha$ is the $N \times K$ matrix with entries $[\boldsymbol{P}_\alpha]_{ij} = \delta_{ij}, \forall ij$, with δ_{ij} denoting the Kronecker delta, and $\boldsymbol{O}$ is the $K \times K$ Haar matrix [18]; ii) iid zero-mean Gaussian matrix ensemble. Note that in the latter case, the sub-optimal variant coincides with the classical AMP recursion [12]. In the former case, with a straightforward calculus we obtain that $\mathrm{S}_{\boldsymbol{A}}(\omega) = (1+\omega)/(1+\omega/\alpha)$ and

$$\lambda_s^t = (1 + \chi^t - \sqrt{(1+\chi^t)^2 - 4\alpha\chi^t})/(2\alpha\sigma_w^2\chi^t) \qquad (50)$$

where $\chi^t \triangleq \left\langle \eta_{t-1}'(\boldsymbol{A}^\dagger \boldsymbol{z}^{t-1} + \boldsymbol{\mu}^{t-1}) \right\rangle / (\alpha\sigma_w^2 \lambda_s^{t-1})$.

In [5, Section 5], the authors report the estimated normalized mean square estimation error (nmsee) of the damped-ADATAP scheme for the scenario (i). Note that for each trial up to 3000 iterations were executed. In Fig. 1 we report the nmsee of the suboptimal variant of S-AMP applied in the same scenario versus the number of iterations. Details are reported in the caption of Fig. 1. Note that no divergence behavior was observed in all performed trials. A comparison of the curves in Fig. 1 with the corresponding curves reported in [5, Fig. 1] shows that both recovery schemes achieve the same performance, but with a significantly smaller number of iterations for the sub-optimal variant.

V. Conclusion

We developed a novel fixed-point algorithm called S-AMP for linear observation models from the equations of the stationary points of the Gibbs free energy under first- and second-moment consistency constraints in the large system limit. AMP is a special case of S-AMP when the measurement matrix has iid zero-mean entries. The optimality of S-AMP follows by its design. We also defined a sub-optimal variant of S-AMP, which is easy to implement. This sub-optimal recovery scheme shows excellent performance when applied in a compressed sensing context to a linear model with a row-orthogonal measurement matrix ensemble. In particular, it converges in around 40 iterations without showing any divergence behavior.

References

[1] A. M. David L. Donoho and A. Montanari, "Message-passing algorithms for compressed sensing," *Proceedings of the National Academy of Sciences*, vol. 106, pp. 18 914–18 919, September 2009.

[2] M. Bayati and A. Montanari, "The dynamics of message passing on dense graphs, with applications to compressed sensing," *IEEE Trans. Information Theory*, vol. 57, pp. 764–785, February 2011.

[3] M. Opper and O. Winther, "Chapter 2-From naive mean field theory to the TAP equations," *Advanced Mean Field Methods: Theory and practice, Neural Information Processing*, 2001.

[4] ——, "Adaptive and self-averaging Thouless-Anderson-Palmer mean field theory for probabilistic modeling," *Physical Review E*, vol. 64, pp. 056 131–(1–14), October 2001.

[5] Y. Kabashima and M. Vehkapera, "Signal recovery using expectation consistent approximation for linear observations," *arvix:1401.5151*, January 2014.

[6] J. S. Yedidia, W. Freeman, and Y. Weiss, "Constructing free-energy approximations and generalized belief propagation algorithms," *IEEE Transactions on Information Theory*, vol. 51, no. 7, pp. 2282–2312, 2005.

[7] E. Riegler, G. E. Kirkelund, C. N. Manchon, M. Badiu, and B. H. Fleury, "Merging belief propagation and the mean field approximation: A free energy approach," *IEEE Transactions on Information Theory*, vol. 59, no. 1, pp. 588–602, 2013.

[8] T. Heskes, W. Wiegerinck, O. Winther, and O. Zoeter, "Approximate inference techniques with expectation constraints," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2005, September 2005.

[9] M. Opper and O. Winther, "Expectation consistent approximate inference," *Journal of Machine Learning Research*, 6 (2005): 2177-2204.

[10] D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing: I. motivation and construction," in *Proc. Information Theory Workshop (ITW)*, Cairo, Egypt, June 2010.

[11] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *Proc. IEEE International Symposium on Information Theory (ISIT)*, Saint-Petersburg, Russia, July 2011.

[12] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová, "Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices," *Journal of Statistical Mechanics: Theory and Experiment*, no. P08009, 2012.

[13] C. Francesco, F. Krzakala, and L. Zdeborová, "On convergence of approximate message passing," *arXiv:1401.6384*, January 2014.

[14] A. M. Tulino and S. Verdú, "Random matrix theory and wireless communications," *Foundations and Trends in Communications and Information Theory*, vol. 1, no. 1, June 2004.

[15] R. R. Müller, G. Alfano, B. M. Zaidel, and R. de Miguel, "Applications of large random matrices in communications engineering," *arXiv:1310.5479*, October 2013.

[16] M. W. Seeger, "Bayesian inference and optimal design for the sparse linear model," *Journal of Machine Learning Research: Theory and Experiment*, vol. 9, pp. 759–813, 2008.

[17] T. Heskes and O. Zoeter, "Expectation propagation for approximate inference in dynamic bayesian networks," in *Proc. Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, 2002.

[18] M. Vehkapera, Y. Kabashima, and S. Chatterjee, "Analysis of regularized LS reconstruction and random matrix ensembles in compressed sensing," *arXiv:1312.0256*, December 2013.